Birdcall retrieval from environmental acoustic recordings using image processing

Xueyan Dong[1], Phil Eichinski[2], Michael Towsey[3], Jinglan Zhang[3], Paul Roe[3]
School of Electrical Engineering and Computer Science
Queensland University of Technology (QUT)
Brisbane, Australia
[1]xueyan.dong@student.qut.edu.au
[2]phil.eichinski@hdr.qut.edu.au
[3]m.towsey, jinglan.zhang, p.roe@qut.edu.au

***Abstract*—Acoustic recordings of the environment provide an effective means to monitor bird species diversity. To facilitate exploration of acoustic recordings, we describe a content-based birdcall retrieval algorithm. A query birdcall is a region of spectrogram bounded by frequency and time. Retrieval depends on a similarity measure derived from the orientation and distribution of spectral ridges. The spectral ridge detection method caters for a broad range of birdcall structures. In this paper, we extend previous work by incorporating a spectrogram scaling step in order to improve the detection of spectral ridges. Compared to an existing approach based on MFCC features, our feature representation achieves better retrieval performance for multiple bird species in noisy recordings.**

Keywords— birdcall retrieval; spectral ridge detection; spectrogram downsampling; environmental acoustic recordings

I. INTRODUCTION

Birds are regarded as good indicator species of environmental health because they provide vital ecosystem services [1]. Automated recording devices are widely used to survey bird species richness because the method works at large spatiotemporal scale and audio-recordings provide a persistent record of the acoustic soundscape in a local region [2-4].

To explore large volumes of environmental acoustic data, many sound analysis algorithms [5] and tools [6] have been developed for automatic call detection and species identification. With careful settings, these tools can be efficient for batch processing and obtain high accuracy on high signal-to-noise ratio recordings. However, their application to noisy recordings, e.g. environmental field recordings where noise is unconstrained, becomes ineffective. Since environmental recordings potentially show advantages for ecological applications such as species presence or absence, more efforts on analysis of those recordings are required.

Automated classification of bird species requires the extraction of discriminative features. Many time domain [7] and frequency domain [8] acoustic features have been investigated. Some features are useful for laboratory recordings or "trophy" recordings which contain the calls of a single species and noise is controlled but they are not suitable for environmental recordings where the calls of multiple species overlap in frequency and/or time. To address this challenge, some researchers have turned to image processing techniques for extracting spectrogram features and have achieved promising results [9, 10].

Another limitation of feature extraction for classification tasks is that the optimum features are task dependent. A particular feature may only be useful for a limited group of species in an acoustically controlled environment. They can behave unexpectedly when applied to recordings with novel content.

Given the limitations of the birdcall classification task, this paper focuses on the *content-based* birdcall *retrieval* task. A requirement of such a system is that its feature set should be more generic and able to describe arbitrary bird call.

In this paper we explore a generic feature descriptor for characterizing a wide range of birdcall structures. The algorithm detects spectral ridges in the spectrograms of audio recordings. A local descriptor is used to describe the distribution and orientation of spectral ridges. This method is compared with an existing approach which uses cepstral coefficients and hidden Markov models (HMMs) on environmental acoustic recordings containing 20 birdcall classes.

II. RELATED WORK

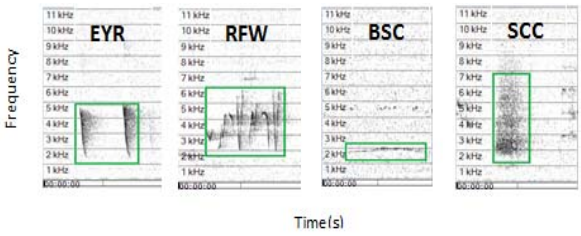
Many studies have been made on content-based audio classification and retrieval in general audio collections [11, 12]. In these retrieval systems, classifiers are used for determining the similarity in term of feature metrics, such as nearest-neighbour [11], support vector machines [12], and HMMs [13]. However, it is choice of features rather than the classifier that is the key to the success of bird species classification and retrieval. Many studies on automatic analysis of bird sound recordings focus on classification system. In these applications, features are often designed for specific purpose. Mel-frequency cepstral coefficients (MFCCs) offer a compact parametric representation of bird sounds with broadband characteristics and harmonics like human speech [14, 15]. However, Somervuo et al. [16] indicated that ceptral coefficients misrepresent important pitch information and they may not be the best choice for bird species. The time-varying sinusoid is

widely used features for characterizing birdcalls consist of modulated tonal sounds [17, 18]. Jančovič and Köküer reported that sinusoidal models can provide a better representation of birdcalls in field recordings than MFCCs [19]. To further explore the use of sinusoidal models, Chen and Maher [20] proposed a method to detect spectral peak tracks by including a variable number of sinusoidal models. Track frequencies, frequency differences, relative power, shape, and duration were extracted as features from detected track to differentiate bird species. The features can describe birdcalls consisting not only of pure tonal components but also harmonic and inharmonic combinations of tones. However, this method is not able to represent calls containing rapidly modulated whistles and clicks (appearing as steep or vertical ridges in a spectrogram) and therefore cannot represent a wide range of birdcalls containing these elements. In a survey of birdcall classification techniques, Brandes [21] reported that most existing features are suitable for some of four basic birdcall components (whistles, chirps, pulses, harmonics) but not for broadband components.

Audio signals are usually converted to spectrograms by applying the short-time Fourier transform (STFT) and are amenable to image processing techniques. Two recent examples are the MPEG angular radial transform [22] and Histograms of Oriented Gradients (HOG) [10]. The features derived from MPEG angular radial transform require that the duration of audio recordings is fixed at three seconds long. The application to varied length of birdsong recordings needs further exploration. HOG features were combined with other acoustic features in the birdcall classification task [10] but the contribution of the HOG features to the final result is not reported separately. Bardeli [23] employed the structure tensor to select "points of interest" and applied a 2D Fourier transform to the neighbourhood of each selected point.

The work we describe in this paper is an extension of our previous investigation using spectral ridges to characterize bird calls in spectrograms [24]. The spectrograms of most bird calls consist of spectral ridges due to whistles (which appear as horizontal ridges in the spectrogram); clicks (which appear as one or more repeated vertical ridges); chirps (which appear as rising or falling ridges); and harmonic tones (appearing as stacks of ridges). Examples are shown in Fig. 1 (a, b, c). However, these intrinsic call features are frequently followed by 'tails' and 'shadows' due to echo in the local environment. Most feature extraction techniques respond to both the intrinsic call and to the echo (an extrinsic feature). Dong et al. [24] report that a feature set derived from spectral ridges (describing only intrinsic bird call features, such as the magnitude, orientation, and distribution of local spectral ridges) yielded more accurate retrieval than comparison methods.

However their method did not work well for three of 19 species whose shriek-like calls had a diffuse structure and lacked clear spectral ridges (see for example Fig. 1 d). In this paper we introduce an additional image processing step, spectrogram scaling, which greatly improves the retrieval performance on birdcalls whose acoustic energy is temporally and spectrally diffuse.



(a) (b) (c) (d)

Fig. 1. Example spectrograms for four birdcalls (from left to right) that are (a) *Eastern Yellow Robin*, (b) *Rufous Whistler*, (c) *Bush Stone-curlew*, and (d) *Sulphur crested-cockatoo*.

III. METHOD

A. Dataset

The dataset consists of 300 one-minute recordings collected from open bushland located in the Samford Ecological Research Facility (SERF) 20 km north-west of Brisbane CBD, Queensland, Australia. The recordings were selected from five days of continuous recordings (13th -17th of October) in 2010 at four sites. The details of the recordings can be found in [4]. The recordings were chosen so as to include a representative range of 20 birdcall classes from 19 species, each class represented by 15 recordings. The selected 20 classes included a range of call structures, both single and multi-syllable calls, with syllables consisting of clicks, chirps, whistles and shrieks. Example spectrograms are displayed in Fig. 1.

Each recording contains the target species but also other acoustic events, such as vocalisations of other animals (dogs, insects, and untargeted birds), traffic noise etc. The recordings for any one species were selected from either different sites or different time of the same day to minimise the probability that calls of the same individual appeared in more than one recording.

The ground truth data is composed of annotations (denoting the temporal and spectral bounds of the call) and labels for all instances of the target call classes present in the dataset. For conducting the proposed algorithm, the 300 recordings were divided into three sets equally, 100 each for query, training, and testing. The selections were made randomly following a rule that each set of recordings includes five recordings representing each of the 20 call-classes.

B. Feature extraction

For feature extraction, we introduce the major steps in the spectral ridge features method: preprocessing, ridge detection, feature representation.

1) Spectrogram preparation: the first step is to prepare a noise reduced spectrogram. Each one-minute recording is formatted with a sampling rate of 22,050 Hz and 16-bit resolution. Spectrograms are generated using the short-time Fourier transform (STFT) with a window of 512 samples and a Hamming window of 50% overlap. We denote spectral values by $X_{(t,\, f)}$, where t represents a window/frame and f indexes a discrete frequency bin. Spectral amplitude values are converted to decibels (dB) using $\mathrm{dB} = 20\log_{10}(X)$. To reduce background noise, we apply an existing noise removal algorithm developed

by Towsey et al. [25] which calculates a separate decibel threshold for each frequency bin assuming an additive noise model. In our application the default threshold is adopted.

2) Spectral ridge detection: we use the ridge detection method in [26] to identify portions of the spectrogram that have ridge characteristics. In the method, ridge points are detected by convolving each spectrogram with four masks, one mask for each ridge direction. Here we employ the set of masks for the directions 0, π/4, π/2, and 3π/4 radians. A pixel in the spectrogram is assigned a ridge direction corresponding to the mask yielding maximum convolution score only if the score exceeds a threshold of 6.0 dB.

The additional processing step introduced in this work is spectrogram scaling consisting of three steps:

Step 1: apply ridge detection on original spectrogram and save the ridge detection results;

Step 2: scale down the spectrogram in the temporal direction by a scale factor (σ_t) and perform vertical ridge detection as previously described. Store the detected points of interest. Repeat the same in the frequency direction with scaling factor σ_f, and store the additional (horizontal ridge) points of interest. Although we are treating the spectrogram as an image, in fact its two dimensions are not spatially equivalent, hence we believed it necessary to scale down each dimension separately.

Step 3: the stored ridges derived from the original and the down-scaled spectrograms are combined to the final ridge output. In detail, the (new) ridges detected from compressed spectrograms are added to the ridges derived from original spectrogram. This process is achieved by expanding at a reverse scale (1/ σ) so as to be the same size as original spectrogram. If a position of new ridge on the spectrogram has no ridge found in previous ridge detection, the new ridges were added.

Fig. 2 shows the process of our ridge detection method. In Fig. 2(a), few ridges can be detected from the original spectrogram while vertical ridges are clearly seen in resized spectrogram in Fig. 2 (b). Notice Fig. 2 only exhibits the effect

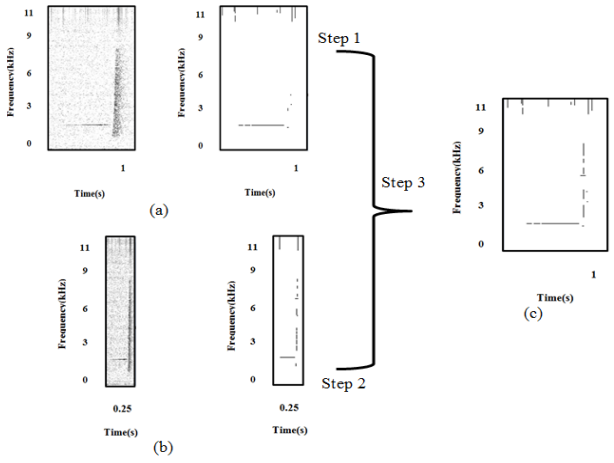


Fig. 2. Spectral ridge detection results showing spectrogram and corresponding detected ridge. Ridge detection on (a) original spectrogram, (b) down scaled spectrogram, (c) final ridge results.

of down-sampling in the temporal direction. Through experiments on training data, σ = 1/4 was chosen for both temporal and frequency compression.

3) Feature representation: the selected birdcalls vary greatly in duration and bandwidth, therefore finding a "universal" feature representation can be hard. We follow the method of normalized block descriptor used in human detection [27] to capture local variations in birdcalls. The call region is divided into a grid of 11×11 non-overlapping blocks. Finally the call is characterized as the vector of all block features within the call region. When searching for potential candidates, the query grid is applied to the one-minute recordings, traversing in steps of five frames. To reduce the computation cost, a region from one-minute spectrogram is eliminated as a matching candidate if 50% of the candidate blocks underlying the query region do not contain ridges. Typically, a one-minute audio file contains about 100 regions that are passed to the next step for feature extraction.

Three spectral ridge features were extracted from each block: 1. temporal entropy; 2. frequency bin entropy; 3. a histogram of four ridge directions.

1. Temporal entropy (H_T): The ridge magnitudes are summed frame-wise over all frames in the block and the N values are normalized to unit sum. H_T is calculated as:

$$H_T = -\sum \mathrm{p_i} \log_2 \mathrm{p_i} \quad \text{where i} \in [1, N] \tag{1}$$

2. Frequency bin entropy (H_B): Similar to the calculation of H_T except that the ridge magnitudes are summed bin-wise over all bins in the block.

3. Histogram of counts of four ridge directions (HoRC4): to further describe the local property of each block, we calculated a histogram of four ridge directions feature that is inspired by Histogram of Oriented Gradient [27]. Here a four-dimensional vector derived from the counts of block pixels belonging to ridges having direction 0, π/4, π/2 or 3π/4. The histogram values are normalized to [0,1].

The entropy features describe the spatial distribution of ridge cells within a block while the HoRC4 feature describes the distribution of ridge directions.

C. Similarity score

Euclidean distance is applied to two feature vectors derived from corresponding blocks. The distance (*d*) is converted to similarity (*s*) using (2). The similarity score (*S*) between a query call and a candidate call is computed using (3).

$$s = 1 - \frac{d}{\sqrt{D}} \tag{2}$$

$$S = \frac{\sum_{i=1}^{n} w.s_i}{n} \tag{3}$$

where *w* is the weights to different blocks, *w* = 0.2 when blocks contain no ridges and w = 0.8 for ridge block.

$$S_{max} = \sum_{i=1}^{n} w \cdot s_i \text{ , where all } s_i = 1.0 \tag{4}$$

Since *n* is dependent on the size of birdcalls, the final *S* is divided by maximum score (*Smax*) in (3) for normalization.

In the spectral ridge method, the idea is to retrieve bird vocalizations from a set of recordings (in this case 100 one-minute recordings) similar to a user-supplied query. The query vocalizations are present in query set. Once a query is provided to our system, our method will search through every one-minute recording to find the location in each file with the highest similarity score. It is important to note that we return only the highest scoring match from each recording. The training data was used to find appropriate parameters for the algorithm that gave the highest retrieval performance. Testing data was only used after parameters had been decided based on training.

D. *Comparative method: Cepstral Coefficients and Hiddern Markov Models(HMM)*

As a comparison, we adopt Song Scope classification algorithm used for identifying vocalizations from continuous recordings. The algorithm extracts spectral features from cepstral coefficients similar to Mel Frequency Cepstral coefficients (MFCCs) to build Hidden Markov Models (HMMs) for recognition.

In the Song Scope algorithm, recordings are first transformed into spectrograms using Fast Fourier Transform (FFT). In preprocessing stage, several techniques are incorporated for noise reduction. Wiener filter is used to remove background noise typically present in field recordings. A band-pass filter removes the unnecessary range of frequencies as birdcalls only occur in fixed frequency range. Finally, power normalization and log frequency scaling is performed for limiting spectral values to a small dynamic range such that noise can be greatly depressed and audio signals can clearly stand out. Next, a simple signal detection algorithm is employed to automatically segment syllables in vocalizations. Calls are then represented by a series of time-related spectral features by applying a series of Discrete Cosine Transform (DCT) on power-normalized and log-warped frequency bins. Finally, HMMs are built to model the features of individual syllables and the syntax of a complex song composed of several syllables. A detailed description of this algorithm is provided by Agranat [5].

The models for the 20 classes are built by taking the manually annotated vocalizations in the training data (100 recordings). The trained models will be applied to test data as a batch, meaning the parameters involved in the recognition are globally set. A good setting of parameters should not only achieve a high sensitivity but also consider the extreme cases of the appropriate values for each classifier. The parameter setting in the experiment can be found in Table I.

TABLE I. PARAMETERS IN SONG SCOPE

Parameters	Values
Sample Rate	22050 samples per second
FFT Window Size	512 samples (32ms)
FFT window overlap	50%
Band-pass filter	500 Hz – 9000 Hz
Max Syllable duration	1.5 seconds
Max Syllable Gap Duration	0.5 seconds
Max Song Duration	3.0 seconds
Dynamic Range	20 dB
Max HMM Model States	48
HMM feature vector size	15
Algorithm	2.0

E. *Evaluation measures*

In this study we assess accuracy at top N results, which is widely used in music retrieval [28]. The accuracy for a given number of queries (in our case, 100) is given by equation (4). In our case, a returned list of C retrieved calls is scored as a correct response to a query if it contains at least one call in the same class as the query.

$$accuracy = \frac{Number\ of\ correct\ retrievals}{Total\ number\ of\ queries} \quad (4)$$

Accuracy by this definition will increase monotonically with the size, C, of the retrieved list.

For simplicity, the query set is not involved in the HMMs-based method. Only training data is used for building models, which are then used on the testing set. For each recording the algorithm only output the instance giving the highest probability score (provided in classification output of Song scope) among the list of identified instances by each model (classifier). Maximum probability score (1.0) indicates the most likely to be the class represented by the model. If the instance has the same label with one provided in the annotation, we count it as a correct retrieval.

Typically, the retrieval results responding to a query are ranked in a descending order of their similarity score. The rank of the first correct retrieval can be used for evaluating the spectral ridge method. We compute the average rank of first correct retrieval over 5 queries within each class.

IV. EXPERIMENTS

The experiments were implemented on previously unseen test data. There are 5 one-minute recordings for each of 20 classes of query call, and each recording contains on average 20 individual vocalizations of that call class as well as other undefined acoustic content.

A. *Experimental results*

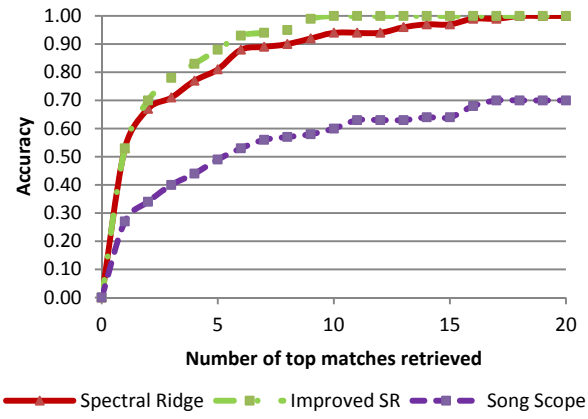


Fig. 3. Retrieval accuracy versus size of the retrieved list.

Fig. 3 shows retrieval accuracy using the previous spectral ridge (SR) method [24], the new scaled version of SR

(Improved SR) and the Song Scope approach. Both SR methods outperform the HMM-based method. In addition scaled-SR outperformed SR. For example, scaled-SR achieved 100% correct retrieval within the top 10 ranked retrievals whereas SR did not achieve 100% correct retrieval until the top 20 ranked retrievals.

TABLE 2. RETRIEVAL PERFORMANCE OVER THE 20 CALL CLASSES FOR THE SPECTRAL RIDGE METHOD ON UNCOMPRESSING SPECTROGRAM (SR) AND ON COMPRESSING SPECTROGRAM (ISR).

Species Common Name	**Text description of call structure**	**Av. rank of first correct retrieval**	
		SR	***Scaled-SR***
Brown Cuckoo-dove	A short chirp	2.2	2.4
Brown Honeyeater	Repeated chirps	2.2	2.2
Bush Stone-curlew	A long whistle	3.6	2.4
Eastern Whipbird	Whistle followed by click	1.0	1.0
Eastern Yellow Robin	Two clicks	2.8	1.8
Grey Fantail	Stack of whistles and clicks	1.0	1.0
Grey Shrike-thrush	Three clicks followed by a whistle	1.2	1.2
Golden Whistler	Broken chirps follow by a chirp	3.0	3.2
Leaden Flycatcher	Stacked broken chirps repeatedly	4.6	4.4
Olive-backed Oriole	Three stacked whistles	2.2	2.4
Rufous Whistler	A long chirp	2.4	2.4
Rainbow Lorikeet	Fairly diffused energy	**12.0**	**6.2**
Shining Bronze-cuckoo	Repeated chirp	2.0	1.2
Sulphur-crested Cockatoo	Shriek	**7.6**	**1.8**
Silvereye	A click follow by a warble	**7.4**	**4.8**
Scarlet Honeyeater (call)	A click follow by a short whistle	1.0	1.2
Scarlet Honeyeater (song)	Multiple chirps	1.2	1.4
Striated Pardalote	Three broken descending whips	2.2	1.2
Torresian Crow	Stacked harmonic	1.0	1.0
White-throated Honeyeater	4 whips repeatedly	3.8	2.2
Average		**3.22**	**2.27**

To further investigate the efficacy of the scaled-SR method, we provide average rank (over five queries) of the first correct retrieval in each call class (Table 2). The three species for which scaled-SR yields significantly increased retrieval are those with diffuse broadband components, *Rainbow Lorikeet*, *Sulphur-crested Cockatoo* and *Silvereye*.

B. Discussion

The results indicate that including a spectrogram scaling step into spectral ridge detection significantly improves the retrieval of bird calls whose spectral energy is diffuse. In addition, scaling also slightly improved the retrieval performance on the 17 species whose call structure was not classed as 'diffuse' (thereby raising their average rank of first correct retrieval from 2.2 to 1.9).

It is likely that the poor performance of Song Scope was as much due to poor syllable segmentation as to the known poor performance of MFCC features in the presence of noise. Song Scope's segmentation algorithm depends on setting appropriate thresholds for signal energy and 'gaps' between syllables. However, as already noted, environmental recordings typically contain many kinds of unwanted acoustic events apart from the species of interest. For example, Song Scope failed to retrieve the test calls of *Shining Bronze-cuckoo* because they overlap with other calls, which does not occur in the training data. This highlights one of the advantages of the retrieval task. The user manually segments the query (thus the segmentation is accurate) but more important, prior segmentation of the candidate calls is not required.

Both SR and scaled-SR perform well in retrieving distinctive call classes, such as *Grey Fantail*, *Eastern Whipbird*, *Torresian Crow* and *Grey Shrike-thrush* because these vocalizations have little variation in structure. In the matching of short calls (such as Scarlet Honeyeater call) and overlapping calls (*Shining Bronze-cuckoo*), nearest-neighbor based on spectral ridge features shows more promising results than Song Scope algorithm. HMMs require more time-varying information than provided by short calls, whereas the SR method handles short-duration queries well.

CONCLUSION

This paper presents a method for retrieving birdcalls from recordings of the natural environment. The method extracts low level image-based spectral ridge features from the spectrograms of audio recordings. The features work well in characterizing and distinguishing various birdcall structures. In contrast to a HMM-based method, it achieves better performance when only a small amount of training data is available. Our new scaled-SR method overcomes a drawback of the previous SR method in that it can now characterize the underlying structure of calls that have diffuse energy. In the future, we plan to investigate the application of our method on a large data set incorporating several months of audio recordings.

REFERENCES

[1] R. D. Gregory, A. Van Strien, P. Vorisek, A. W. G. Meyling, D. G. Noble, R. P. Foppen, and D. W. Gibbons, "Developing indicators for European birds," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 360, pp. 269-288, 2005.

[2] J. Wimmer, M. Towsey, B. Planitz, P. Roe, and I. Williamson, "Scaling acoustic data analysis through collaboration and automation," in 2010 IEEE Sixth International Conference on e-Science (e-Science), 2010, pp. 308-315.

[3] K.-H. Frommolt, K.-H. Tauchert, and M. Koch, "Advantages and disadvantages of acoustic monitoring of birds–realistic scenarios for automated bioacoustic monitoring in a densely populated region," Computational Bioacoustics for Assessing Biodiversity. Proc. of the Internat. Expert Meeting on IT-based Detection of Bioacoustical Patterns. BfN-Skripten, vol. 234, pp. 83-92, 2008.

[4] J. Wimmer, M. Towsey, P. Roe, and I. Williamson, "Sampling environmental acoustic recordings to determine bird species richness," Ecological Applications, vol. 23, pp. 1419-1428, 2013.

[5] I. Agranat, "Automatically identifying animal species from their vocalizations," in Fifth International Conference on Bio-Acoustics, Holywell Park, 2009.

[6] I. WildlifeAcoustics. (2015, January 21, 2015). Song Scope Software. Available: http://www.wildlifeacoustics.com/products/song-scope-overview

[7] E. D. Chesmore, "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," Applied Acoustics, vol. 62, pp. 1359-1374, 2001.

[8] T. S. Brandes, "Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, pp. 1173-1180, 2008.

[9] Brandes, "Techniques for bioacoustic signal detection using image processing," Computational bioacoustics for assessing biodiversity. Bundesamt fur Naturschutz, Bonn, Germany, pp. 103-110, 2008.

[10] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," The Journal of the Acoustical Society of America, vol. 131, pp. 4640-4650, 2012.

[11] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," MultiMedia, IEEE, vol. 3, pp. 27-36, 1996.

[12] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," IEEE Transactions on Neural Networks, , vol. 14, pp. 209-215, 2003.

[13] H.-G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representations," IEEE Transactions on Circuits and Systems for Video Technology,, vol. 14, pp. 716-725, 2004.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, pp. 357-366, 1980.

[15] J. W. Picone, "Signal modeling techniques in speech recognition," Proceedings of the IEEE, vol. 81, pp. 1215-1247, 1993.

[16] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 2252-2263, 2006.

[17] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04), 2004, pp. V-701-4 vol. 5.

[18] P. Jančovič, M. Köküer, and M. Russell, "Bird species recognition from field recordings using HMM-based modelling of frequency tracks," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 8252-8256.

[19] P. Jančovič and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," EURASIP Journal on Advances in Signal Processing, vol. 2011, p. 982936, 2011.

[20] Z. Chen and R. C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," The Journal of the Acoustical Society of America, vol. 120, pp. 2974-2984, 2006.

[21] Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," Bird Conservation International, vol. 18, pp. S163-S173, 2008.

[22] C.-H. Lee, S.-B. Hsu, J.-L. Shih, and C.-H. Chou, "Continuous Birdsong Recognition Using Gaussian Mixture Modeling of Image Shape Features," IEEE Transactions on Multimedia, vol. 15, pp. 454-464, 2013.

[23] R. Bardeli, "Similarity search in animal sound databases," Multimedia, IEEE Transactions on, vol. 11, pp. 68-76, 2009.

[24] X. Dong, M. Towsey, A. Truskinger, M. Cottman-Fields, J. Zhang, and P. Roe, "Similarity-based birdcall retrieval from environmental audio," Ecological Informatics, 2015.

[25] M. Towsey, J. Wimmer, I. Williamson, and P. Roe, "The use of acoustic indices to determine avian species richness in audio-recordings of the environment," Ecological Informatics, vol. 21, pp. 110-119, 2014.

[26] X. Dong, M. Towsey, J. Zhang, J. Banks, and P. Roe, "A Novel Representation of Bioacoustic Events for Content-Based Search in Field Audio Data," in Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on, 2013, pp. 1-6.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, 2005, pp. 886-893.

[28] J.-S. R. Jang and H.-R. Lee, "Hierarchical filtering method for content-based music retrieval via acoustic input," in Proceedings of the ninth ACM international conference on Multimedia, 2001, pp. 401-410.